Automated Fact-Checking from Scientific Research: A Systematic Review of the State-of-the-Art Methods

Salmane Dazine
Vrije Universiteit Amsterdam, The Netherlands
s.dazine@student.vu.nl

ABSTRACT

Fact-checking is a challenging task because verifying claims is highly dependent on the evidence relevant to the claim. Mitigating this challenge by selecting reputable scientific research papers as evidence implicates presenting a new challenge of investigating, selecting, and retrieving the correct evidence from a wide pool of research. In this study, I present the overall pipeline employed in automated fact-checking, I list 7 datasets that can be used to train scientific fact-checking models, and I review 19 methods that can be used in the task of fact-checking textual claims with the presence of evidential textual documents i.e. scientific research papers. After the review, I compare and evaluate all the methods and suggest the best techniques for dataset training, evidence retrieval, and veracity prediction that proved to be successful based on their reporting studies. Confirming or falsifying these suggestions requires further experimental work in the thesis project which presents the future work of this study.

KEYWORDS

Systematic Literature Review, State-of-the-Art, Algorithm, Model, Classifier, Training, Testing, Neural Network, Data-set, Automated Fact-checking, Truth, Truthfulness, Veracity, Credibility, Statement, Claim, Fake, Check-worthy, Detection, Document, Evidence, Retrieval, Justification, Explanation, Verdict, Rationale, Validation, Verification, Knowledge Base, Knowledge Graph, Rule Mining, Web, Social Media, Academic, Research, Papers, Scientific, Literature

1 INTRODUCTION

With the rise of online content circulating through social media, news channels, and various kinds of online platforms, the potential for false information continues to grow. Non-factual information could be seen as disinformation which is intentionally meant for spreading incorrect claims. Still, it could also be misinformation that is not necessarily intended to be harmful [17]. In both cases, information pollution could lead to destructive effects on society. After COVID-19, the World Economic Forum considered misinformation to be one of the factors that impeded the response to the pandemic, and that its threat will be heightened in the future [55]. Although it might be true that fake news could come from non-authoritative sources, like online forums or non-reputable social media users, a study reported by X. Li et Al. [30] that measured the accuracy and coverage of news sources in the stock and flight domain, proved that even authoritative sources don't always provide accurate news. 86% for the stock domain and 80% for the flight domain were the respective measured average accuracies of the claims provided by the sources. Potential copying between sources was detected as a common factor for spreading false news. In addition to inconsistent values that were explained by: semantics ambiguity, instance ambiguity, out-of-date data, and pure erroneous data. Regardless of its causes, spreading false information was met by a growing interest in automated fact-checking. Many works succeeded in assessing the credibility of textual claims based on the reliability of the sources and the context and discourse features where it is found. However, less interest is given to well-cited and peer-reviewed research papers as evidence. Predicting the truthfulness of claims by using evidence from web articles or news websites might defeat the overall purpose of fact-checking. Source reliability is an important factor that also affects the veracity of claims. By resorting to peer-reviewed and well-cited scientific research papers, the risk of using false evidence is mitigated. This, naturally, comes with the drawback of finding, understanding, and selecting the correct evidence from a large list of papers. Using scientific papers is particularly challenging because it eradicates the assistance of structured knowledge bases, and requires creating a new knowledge base by using natural language processing tasks to convert unstructured data to structured data. In this paper, I categorize, summarize, and assess the state-of-the-art methods that can be utilized to verify the credibility of textual claims by retrieving the evidence from scientific research papers. After presenting a general overview of the pipeline of automated fact-checking, I classify the proposed algorithms based on their focus, and I summarize their main attributes, strengths, and weaknesses. Finally, I discuss their applicability and efficiency to verify statements by resorting to scientific research.

2 RELATED WORK

Present works are dominated by general fact-checking surveys. These include the surveys by Y. Li et al. [31], Z. Guo et al. [17], X. Zeng et al. [71], G. Karadzhov et al. [23], J. Thorne et al. [62], and N. Kotonya et al. [25]. Nevertheless, there is no published survey or review that scopes down the fact-checking methods to the scientific field. This study fits in the overall research picture by focusing on methods that can be either directly used in the scientific scope, or indirectly used by employing some of their techniques to fact-check textual statements by resorting to a collection of evidential textual documents. Although there are no secondary studies that combine all of the scientific fact-checking methods, there is a large pool of primary studies that were used to build this systematic review. The methods that can be used directly in the scientific scope are mainly those that came as a presentation or submission to the Sci-Fact challenge. Those include the papers by R. Pradeep et al. [51], X. Zeng et al. [72], X. Li et al. [29], D. Wadden et al. [66], Z. Liu et al. [34], and N. Kotonya et al. [25]. The rest of the papers indirectly contribute to the topic by providing general-purpose methods that can still be applied in the context of fact-checking textual claims from scientific research papers. These consist of all the other methods that are not related to the Sci-Fact challenge.

3 STUDY DESIGN

3.1 Research Question

What are the state-of-the-art methods that can be utilized to detect the credibility of textual statements by resorting to evidence from scientific research?

3.2 Research Goals

The research goals of this study can be formalized as follows:

- Reporting comprehensively the current state-of-the-art methods for automatically fact-checking textual statements by resorting to evidence from scientific research.
- Evaluating and comparing the methods learned from research.
- Learning to distinguish models suitable for different fact-checking tasks, primarily based on the type of the claim and the evidence.
- Facilitating an overall methodology to be followed for solving a real-life scientific fact-checking project.

3.3 Initial search

The initial search for literature was carried out as follows:

- Studying the Fact Checking lecture from the Web Data Processing Systems course [64] and including the papers cited in the lecture in the initial list of selected studies.
- Translating the formulated research question to a search query based on its keywords and feeding it to Google Scholar.
- Selecting an initial list of 10 studies that fulfill all inclusion criteria and is clear from all exclusion criteria.
- Beyond the criterion, a good rationale for selection in the initial list is the comprehensiveness of the study. In order to increase familiarity and to well situate the topic, priority was given to secondary studies that include literature surveys and systematic reviews, and then to primary studies consisting of novel technical contributions that target specific fact-checking problems in detail.
- After reading secondary studies, priority was later given to primary studies that provide the desired level of detail in explaining the methods.

3.4 Application of selection criteria

The selected studies in addition to their inclusion/exclusion criterion were tracked using the attached "Selected Literature" excel file. The selection criterion in terms of inclusion and exclusion are as follows:

I1- Studies written in English.
I2- Studies that are peer-reviewed and well-cited.
I3- Studies proposing reusable and applicable fact-checking systems in the context of the time and technology resources available for a thesis project.
I4- Studies clearly and concisely explaining all the steps, and technologies applied in their proposed fact-checking systems.
E1- Studies that rely on sources of evidence or datasets that are not publicly accessible.
E2- Studies that utilize outdated methods or technologies.
E3- Studies in the form of web articles, short papers, and tutorials that do not provide a sufficient level of detail.

3.5 Snowballing

The Snowballing approach was used relatively more frequently in secondary studies like literature surveys and systematic reviews. These studies categorize fact-checking methods based on different factors. This makes it easier to navigate and distinguish the methods and to select further papers that provide more details about the models and algorithms employed to solve the fact-checking problem. Papers that compare a list of fact-checking methods open a portal to a new list of papers to read to understand the methods addressed more thoroughly.

3.6 Data Extraction

In order to gather the data for the study, two different data extraction processes have been utilized based on the type of the study.

3.6.1 Primary Studies. For systematic reviews and literature surveys, data has been extracted from all the relevant sections of the paper. By summarizing the whole paper, we can find data about related works, the fact-checking pipeline, the training datasets, the categories of the methods employed, the challenges encountered in the field, etc.

3.6.2 Secondary Studies. For technical projects, the data extracted consisted of a summary of the models and algorithms proposed to solve a specific task. The data needed from the methods utilized consisted of:

- A definition of the fact-checking task
- A summary of the model utilized to solve the task
- A subjective evaluation of the model described

3.7 Data Synthesis

After extracting the data, the next step is to synthesize it. For each paper, the data extracted is categorized based on the section that it could potentially belong to in this review e.g. introduction, methodology, results, etc. The data is then carried from its main source to its relevant section. Combining data from different sources in one section could reveal repetitions or discrepancies. Everything is highlighted, synthesized, and cited in a coherent and narrative text.

4 FACT CHECKING OVERVIEW

4.1 Definition

Fact-checking a claim simply means assessing whether the claim is true or false [17]. In the Computer Science field, it consists of determining the degree of correctness of a statement by inspecting external data sources for evidence and aggregating and evaluating the evidence collected. The terms Truth Discovery, Fake News Detection, Stance Classification, Veracity Prediction, and Incongruent Headline Detection have all been used interchangeably in different studies to mean a relatively similar concept [31]. However, Thorne et. al. [62] distinguished between fact-checking which addresses the logic of the statement, and fact verification which means getting the right facts by assessing the source and context of the fact.

Despite different studies implementing different fact-checkers, they mostly share the common grounds of being accurate so that predictions mimic the truth, scalable with the least human intervention, and finally, interpretable so that the users understand the rationale behind the predictions.

4.2 Pipeline

The fact-checking pipeline is a series of tasks that are applied in sequence in order to predict the truthfulness of a claim. Methods differ in their approach of dividing the fact-checking operation into a list of individual tasks. Nevertheless, the pipeline could be generalized to the sequence of tasks demonstrated in Figure 1 as suggested by Z. Guo et al. [17]

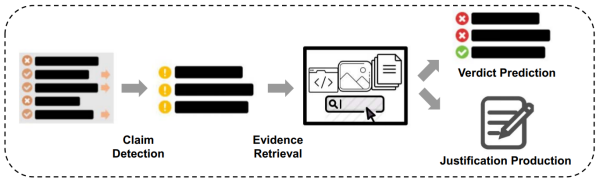

Figure 1: Automated Fact-Checking Pipeline

4.3 Claim

4.3.1 Defintion. Konstantinovskiy's definition of a claim is: "an assertion about a world that is verifiable with readily available evidence" [24]. A claim could be found under different formats [62]:

- Subject-Predicate-Object triples: e.g. (Rabat, capital_of, Morocco). They are popular because they make it possible to use structured knowledge bases, without needing an extra natural language understanding task to convert the relevant text to a triple. A drawback of triples, however, is that they restrict adding more information that could be found in a text.
- Textual Claims: which can be numerical, e.g. "I was born in 1998." Or entity and event properties, e.g. "Morocco is part of the 2022 World Cup." Or position statements, e.g. "Obama supports abortion.", Or quote verifications, e.g. "Trump declared he will run for the presidency."
- Entire Documents: This needs an extra step to extract the relevant textual claims through claim detection.

4.3.2 Claim Detection. Claim detection consists of detecting the claims that contain facts worthy of checking. There have been numerous works about identifying check-worthy claims and rumourous claims, although that can be subjective, and time-dependent [17]. There exist also datasets to help train models to predict check-worthy claims with different sizes, languages, and label classifications [71].
Claimbuster is a system provided through a user interface where the users can input the texts to be checked and receive the fact-checking report [21]. It contains a Claimspotter element that solves a supervised learning problem of detecting check-worthy sentences by constructing a labeled dataset of sentences quoted by presidents during the presidential debates from 1960 to 2012, either as non-factual sentences, unimportant factual sentences, or check-worthy factual sentences. The dataset has been classified using a data collection interface given to recruited participants. And the features used included: the sentiment score, the length, the words included, the part of speech tags, and the entity types [20]. Checkworthiness could also be measured with a scoring system where the higher the score the more likely the claim is check-worthy.
Another study by P. Atanasova et al. [6] worked on check-worthy claim identification of dialog-style texts by analyzing a dataset of political debates CW-USPD-2016 [5]. They used context-based features related to the position of the sentence, segment size, metadata, topic, word embeddings, contradictions, discourse, and similarity of the sentence to other known examples. They have also reimplemented features used in Claimbuster in order to compare their model to the previous state-of-the-art. Afterward, they trained a feed-forward neural network to classify the sentences as positive if one or more sources from nine independent media organizations have fact-checked a claim included in the sentence, and negative otherwise. Then they ranked the sentences based on the scores achieved.

4.3.3 Claim Matching. Claim matching aims to check whether the claim has already been solved by another fact-checker to prevent re-checking it. The ClaimBuster system also includes a Claim matcher that, given the claim spotted, examines in a fact-checking repository for the matching facts based on the similarity of tokens and semantic similarity [20].

4.4 Evidence

4.4.1 Defintion. In order to predict the truthfulness of a claim, a piece of evidence that supports or refutes the claim is needed. This evidence can be found in different formats [62]:

- Graph Data: the most prominent source of this format is Knowledge Graphs or Knowledge Bases, from which it is possible to retrieve the triples corresponding to the claim. J. Thorne et al. [61] provided their framework for verifying numerical claims using knowledge bases, where they find a matching entry in the knowledge base to support or refute the studied claim by: linking named entities in the claim to the knowledge base entities, then extracting all the tuples including the entity, then filtering the non-relevant tuples by identifying the matching relations using a trained binary classifier, and finally comparing the values between the triple and the claim to conclude the label. The drawback of knowledge bases is that they might not be big enough to include rules to support or refute claims.
- Unstructured Data: i.e. web pages, article headlines, documents, social network posts, aggregate information on the distribution of social media posts, metadata of the claim, publication date, user profile, etc. The drawback of unstructured data is text bias. An additional process of natural language understanding and processing is needed before converting this data to understandable evidence [31].
- Structured Data: i.e. web tables, databases, etc. An advantage of structured data is that it comes with more certainty compared to unstructured data, but it may not provide as much information [31].

It is preferred to combine one or more of these formats to achieve better coverage of pieces of evidence [17].

4.4.2 Evidence Retrieval. Evidence retrieval means finding information beyond the claim [17]. This consists of inspecting external sources and identifying the evidence that matches the studied claim. It may include [71]:

- Document Retrieval: this could be performed by querying a search engine using keywords from the claim like named entities, noun phrases, and capitalized expressions, and selecting the most similar documents to the keywords through supervised learning.
- Rationale Selection: it could be performed through one or a combination of keyword matching, sentence similarity scoring, and supervised ranking.

4.4.3 Source of Evidence. In addition to assessing the evidence itself, many studies also considered the source of the evidence to judge the reliability of its provided facts [31]. While some methods assume that sources give their information independently without copying, others try to detect copying relationships based on the similarity of their mistakes. In that sense, they treat the sources of evidence as a group of attributes, with different reliabilities for each attribute. Initially, the reliability of sources could be equal, however, by providing a false claim, the reliability score of these sources takes a penalty. Other than truthfulness, other metrics that can affect the reliability of the sources include coverage or completeness, bias or subjectivity, freshness or updating, and copying.

4.5 Verdict

4.5.1 Defintion. The verdict determines the veracity of the claim [17]. It presents the end result of the fact-checking pipeline. Assessing the truthfulness of the input claim can be done in different formats [62].

- Binary Classification: e.g. True/False. Which is the simplest way to support or refute the claim
- Multi-label Classification: e.g. True, False, Lack of Info, Non-Factual. Recognizes that the claim can not always be classified as either True or False and provides more information about the state of the claim or the evidence extracted.
- Multi-Point Scale: e.g. a range from 0 to 1. This provides more flexibility for ranking the claims based on their trustworthiness and doesn't assume that there is only one true value by rejecting all the other candidates [31].

4.5.2 Verdict Prediction. Verdict prediction consists of verifying the studied claim based on the extracted evidence. In their survey, X. Zeng et al. [71] classified verdict prediction tasks into three main categories:

- Text Classification: deciding whether the claim and the evidence agree, contradict, or if there is a lack of information. More labels could also be added, although that would make it harder to classify the claims.
- Recognizing Textual Entailment: deciding whether the meaning of one text can entail the meaning of another text (support or contradict hypothesis). Machine learning models perform better in this category.
- Natural language Inference: deciding whether a hypothesis can be inferred from a premise. In this category, large neural network models with natural language inference training datasets have been employed.

4.6 Justification

4.6.1 Definition. Justifications or Explanations consist of delivering a rationale for the verdict provided. N. Kotonya et al. [25] distinguished between explainability and interpretability.
The latter is understood as the ability of the model to demonstrate the decision-making process. That is not necessarily understandable by people with no technical expertise. Decision-making processes can be designed to be humanely interpretable. Not only would they show which pieces of evidence were used to give the verdict, but also show the reasoning process before reaching the verdict [17].

4.6.2 Justification Production. Producing explanations means listing the set of rules or sentences that were used to arrive at the final verdict. This could be achieved by summarizing the context of the retrieved rationale, using attention weights to highlight the most important parts of the evidence, or also by using surface-level linguistic features and metadata as part of the justification process [62].
A good justification doesn't always equal a good verdict. Justifications should be readable, plausible i.e. convincing, and faithful i.e. reflecting what happened in the reasoning process [17].

4.7 Challenges

Automated fact-checking is met with various challenges. Different surveys have mentioned different challenges [31][17][71]. The most commonly listed ones include:

- Conflicting Evidence: with the sparsity of sources of evidence, it is likely that the pieces of evidence would conflict. Majority voting is one proposed approach to overcome this. However, it is not always accurate, as the majority of sources could be wrong. Averaging the sources of evidence is another approach, but both of these aforementioned approaches assume that the selected sources are all reliable. As was aforementioned in the introduction, not even authoritative sources may not be 100% accurate. Therefore, source reliability is unknown initially, and has to be inferred. Data sources that offer correct facts more frequently should be given higher reliability scores. Thus, if the claim is supported by high-reliability sources, it has a high chance of being trustworthy.
- Text understanding: whether it is about understanding the input claim, or understanding the evidence sentences. Data can have different meanings. Semantic ambiguity refers to data having different meanings, while instance ambiguity refers to the same name relating to different instances. These challenges could be mitigated through natural language processing algorithms. Further, information comes in different formats, types, and languages. All of these are

issues that need to be considered when dealing with non-structured forms of data.
- Datasets: training and testing fact-checking models is influenced by the quality of the utilized datasets. Dataset imbalance hinders the ability to apply these models to facts from all domains. Only a few domains, like health and politics, dominate claim detection datasets. The quality of the datasets also impedes the accuracy of the models, as released datasets are full of artifacts and biases.

5 ANALYSIS OF LITERATURE

The analyzed literature includes a wide range of methods applied for automated fact-checking. In this report, I include the methods that can be applied in a scientific setting. That is, given a textual claim that could be found in an open setting, or could be given as input by a user, the evidence should be retrieved from textual documents consisting of scientific academic research found online or saved in a local database. Methods that apply to inspecting evidence from academic research papers, whether be it abstracts, full documents, or metadata about the publication are included. Other methods that don't apply to this setting have been omitted.

5.1 Datasets

Before presenting the fact-checking methods, I introduce a list of common datasets that could be employed in training and testing fact-checking models in the research and scientific field. There is a variety of publicly available datasets that differ in the type of input claim, evidence, verdict classification, and languages included. Some can support training and some are not big enough to support it. Regardless of the size of the datasets, the lack of variety in their content makes it difficult to use them for training general-purpose models. Therefore, reducing the scope of the models to scientific claims mitigates the training challenge. Table 1 summarizes the main attributes of the discussed datasets.

5.1.1 *SciFact [67].*

Definition. According to D. Wadden et al., a scientific claim is "an atomic verifiable statement expressing a finding about one aspect of a scientific entity or process, which can be verified from a single source" [67].

Dataset Generation. Sci-Fact is a publicly available dataset consisting of 1409 expert-written scientific claims curated by biomedical experts, annotated with credibility labels, and abstracts containing the rationale behind the label. The verification of the claims has been applied against 5,183 abstracts. Given a scientific claim, and a list of abstracts, the abstracts may either support, refute or provide no information to verify the claim. The rationale is the collection of sentences that are used to predict the label of the claim. While UKP Snopes [18] extracted its claims from political fact-checking websites, and Fever [63] synthetically produced its claims by mutating sentences from Wikipedia articles, the claims in SciFact are retrieved from scientific articles which makes it suitable for training scientific fact-checking models.

Baseline Model. The proposed baseline model is called VeriSci. Following the BERT model [60], VeriSci is a pipeline of 3 components:

- Abstract Retrieval: which retrieves the top k nearest abstracts using the TF-IDF similarity to the claim, inspired by the DrQA TF-IDF implementation by Chen et al. [14].
- Rationale Selection: which detects the sentences containing the evidence from each abstract.
- Label Prediction: using a trained BERT model on all of the FEVER, UKP Snopes, and SciFact datasets, to predict the label of the claim given each rationale sentence.

Evaluation. The submitted models are verified on two levels:

- Abstract Level Evaluation: to assess if the abstract contains evidence and if the predicted label is correct.
- Sentence level Evaluation: to assess if the evidence-holding sentences are identified correctly, and don't include extra rationale sentences.

5.1.2 *SciFact-Open [65].*

Dataset Generation. This new version of SciFact raises the question of scalability and attempts to approximate how fact-checking would be performed in a real-life setting. The performance of 5 of the most accurate state-of-the-art systems developed for SciFact dropped by 15% to 30% on the F1 scale against open domain settings consisting of hundreds of thousands of evidence documents. By extending the verification to a corpus of 500K research abstracts and reducing the number of claims from 1409 to 279. The new corpus extended the original corpus by filtering the S2ORC dataset [36] for all the articles that discuss biology or medicine topics and that have at least an outbound and an inbound citation. This list of articles was then randomly sampled while limiting the corpus to 500K abstracts. To collect evidence from the corpus, an information retrieval system was utilized by combining a BM25 ranking function and a neural ranker based on VERT5ERINI [51]. The system retrieves the nearest abstracts after ranking them based on the confidence scores that they contain evidence. The final labels were produced by two expert human annotators.

5.1.3 *FEVER [63].*

Dataset Generation. FEVER is a publicly available dataset, consisting of 185,445 claims. The claims were generated by annotators by copying or mutating information extracted from Wikipedia. The annotators, then, labeled each claim as Supported, or Refuted and accompanied their classification by the related evidence. Otherwise, if they're not certain, they would label it as NotEnoughInfo. Although the domain of this dataset is not thoroughly scientific, it can still be beneficial in training models related to extracting evidence from textual documents similar to Wikipedia documents.

Baseline Model. The baseline approach presented in the paper consists of:

- Document Retrieval: was also accomplished using the DrQA TF-IDF implementation which returns the top 5 nearest documents.

Dataset	Domain	Claim	Size	Evidence	Source	Verdict	Labels
SciFact	Biomedical	Textual	1,409	Unstructured	Abstracts	Multi-label	3
SciFact-Open	Biomedical	Textual	279	Unstructured	Abstracts	Multi-label	3
FEVER	General	Textual	185,445	Unstructured	Wikipedia introductions	Multi-label	2
MultiFC	General	Textual	34,918	Unstructured	Fact-checking websites	Multi-label	2-27
HealthVer	COVID-19	Textual	14,330	Unstructured	Bing search engine	Multi-label	3
PubHealth	Health	Textual	11,832	Unstructured	Fact-checking and news websites	Multi-label	4
COVID-Fact	COVID-19	Textual	4,086	Unstructured	Google search engine	Multi-label	2

Table 1: Scientific Datasets Attributes

- Sentence Retrieval: using the same method, the top 5 most similar sentences from the documents are extracted using TF-IDF vector similarity.
- Textual Entailment Prediction: between the claim and the evidence passage using a decomposable attention model inspired by A. Parikh et al. [47]

Evaluation. Evaluation is applied by assessing the evidence provided in the cases of supporting or refuting the claims. The correctness of the evidence is evaluated using an F1 score comparing the given evidence sentences to the annotators' evidence sentences. All in all, the Fever score measures the percentage of correct retrieved evidence for supported and refuted categories.

5.1.4 *MultiFC [8].*

Dataset Generation. MultiFC is a publicly available dataset of naturally occurring claims and rich metadata collected from 26 fact-checking websites in English. The list of websites was crawled from Duke Reports' Lab [2] and Wikipedia's Fact-Checking page [3] and further filtered based on the availability of veracity labels and needed types of metadata. The detected claims were also filtered based on duplication and occurrence frequency. Given that they were extracted from a variety of domains, the claims had labeling systems ranging from 2 labels to 27 and were not mapped onto the same scale.

Baseline Model. Two baseline models were presented.

- Multi-Domain Claim Veracity Prediction with Disparate Label Spaces: a multi-task learning model learns how the varying labels can be mapped to one another. Each website domain is modeled as a separate task, and labels are presented in an embedding space. Predictions are calculated through a label embedding layer that takes the dot product between the label embedding and the claim-evidence hidden representations.
- Joint Evidence Ranking and Claim Veracity Prediction: this model considers evidence on top of claim surface patterns. Snippets are extracted from 10 evidence pages and encoded along with their claims to obtain a sentence embedding. Claim-evidence pairs are then ranked and final predictions are calculated through a dot product between label embeddings and ranking scores.

Evaluation. The evidence-based model outperforms the claim-only model in terms of Micro and Macro F1, by a margin of 7% and 4.4% respectively.

5.1.5 *HealthVer [58].*

Dataset Generation. HealthVer is a dataset for evidence-based fact-checking of real-world health-related claims. 14,330 claims were extracted from a search engine that was given questions about COVID-19. After automatically retrieving related articles and manually verifying evidence from their abstracts, the claims were manually annotated as either support, refute, or neutral. The paper argues that training models on real word medical claims outperforms training on synthetic claims.

Baseline Model. Claims accompanied by their evidence were concatenated. The pairs were then fed to BERT [60], SciBERT [9], BioBERT [28], and T5 [52] models that were all trained to make the labeling decision.

Evaluation. The T5 model outperformed the BERT-based models on all metrics of precision, recall, F1, and accuracy. Training on SciFact and HealthVer datasets also proved to achieve better results compared to FEVER although the latter is much larger.

5.1.6 *PubHealth [26].*

Dataset Generation. the dataset contains 11.8k claims related to biomedical subjects, government healthcare policies, and other public health-related topics collected from news and fact-checking websites. The claims are also accompanied by crafted explanations to support the labels consisting of the news summary or the given fact-checking justification. The veracity labels were standardized into a multi-label system of true, false, mixture, and unproven.

Baseline Model. The baseline model presented in the paper consists of the following:

- Classifier for Veracity Production: predicts the probability of the label of the claim using a trained Sentence-BERT model [54].
- Summarization Model for Generating Explanations: employs a joint extractive-abstractive summarization [32] to extract the explanation from the article text of the claim.

Evaluation. For the predictor, four BERT-based models scored better than the baseline model in terms of precision, recall, F1, and

accuracy. While the explainer outperformed a LEAD-3 baseline and a summarization-based ORACLE upper bound.

5.1.7 *COVID-Fact [56].*

Dataset Generation. a FEVER-like dataset containing 4,086 claims related to the COVID-19 pandemic and their relevant evidence. 1,296 claims are extracted from the subreddit r/COVID19 and filtered based on form and trustworthiness. 2,790 contradictory claims refuted by the evidence are automatically generated and included in the dataset by replacing salient words with their opposite. Evidence selection is based on the work of Hidey et al. [22] where they employ cosine similarity on SBERT sentence embeddings to extract the nearest 5 sentences.

Baseline Model. The baseline model presented in the paper consists of the following:

- Evidence Retrieval: the same approach for constructing the dataset was utilized with the help of Google Search.
- Veracity Prediction: evidence sentence embeddings are concatenated and fed to a binary classifier to produce the prediction.

Evaluation. The baseline achieves an F1 score of 32 compared to the FEVER baseline system scoring 18.26. Using COVID-Fact for training also outperformed FEVER by a margin of 37% and SciFact by 25% on the F1 score.

5.2 Methods Categorization

The categorization of automated fact-checking methods was laid out differently between several surveys. J. Thorne et al. [62], for example, classified these methods into supervised models, recognizing textual entailment-based models, and matching claims with existing fact-checked ones. While others [71] [17] composed the methods by the individual task of the fact-checking pipeline they solve. In this survey, I present the different methods of fact-checking using scientific research based on the focus of the method. While some methods focus on the surface form of the claim, the context, and the discourse where it was found, or the metadata of its publication. Others ignore these features and focus solely on finding evidence from websites of research papers and evaluating the reliability of these sources.

5.3 Claim Oriented Methods

Assuming that the textual scientific claim is found online. Incorporating useful information like the context of the claim, the author of the article, and the reliability of the website publishing the article can all be valuable features that help predict the veracity of the claim. For this reason, claim-oriented methods can contribute to fact-checking scientific textual claims. Claim-oriented methods could be one or a combination of supervised classifications with claim-related features, or neural network methods based on sequence modeling [17]. Regardless of the type of modeling utilized, I classify the relevant methods based on the source where the claims are found.

5.3.1 *Answering Questions Forums.*

In Q&A forums, the claims are formatted as answers to the questions posted by the users. Generally, this enables fact-checking methods to resort to the user, thread, and context of the answers to assess their credibility without needing additional evidence.

T. Mihaylova et al. [40].

- Task: SemEval is an annual research workshop aiming to advance state-of-the-art natural language processing methods. This study came as part of Task 8 from 2019 which aims to classify questions and answers from the CQA-QL-2016 dataset from the SemEval-2016 Task3 [43] which contains questions and answers extracted from the Qatar Living forum. The questions have been annotated to factual, opinion, and socializing. While the answers have been annotated to true, false, partially true, conditionally true, responder unsure, and nonfactual.
- Model: in order to classify the answers, the contexts of the answers have been modeled with regard to the entire answer that includes it. The features were based on the similarity to the other answers given, the online threads, and the high-quality posts in the Qatar Living forum, in addition to other discourse and context features. The model combined LSTM-based neural networks with kernel-based support vector machines. The word embeddings are separately fed into a bi-LSTM recurrent neural network to train feature representations for the sentences. These representations given by each text source are then fed into a kernel-based SVM. They are then concatenated and connected to a hidden layer, that outputs the task-specific embeddings, which are finally connected to a softmax output that classifies the claim.
- Evaluation: When it comes to the baseline model, the accuracy of the question classification achieved an accuracy of 45% while the answers classification achieved an accuracy of 83% on all information sources.

P. Nokav et al. [44].

- Task: using the same SemEval-2016 data [43] from the Qatar Living Forum, questions with at least 10 answers have been selected and labeled based on goodness and credibility, in order to create a new corpus.
- Model: the selected features modeled the user, the answer, the question, the thread, and the interaction between them. An SVM rank learning algorithm with linear and RBF kernels has been employed.
- Evaluation: evaluation consisted of experimenting with features separately and also with combinations of the best k features to evaluate their accuracy. The trollness feature [39] achieved the highest mean average precision, average recall, and mean reciprocal rank, among all features. While the top-6 combinations of features resulted in a further improvement in the MAP and AvgRec scores.

5.3.2 *Social Network Posts.*

Claims found in social media work similarly to Q&A forums. However, they may include more important features relating to the propagation and circulation of the posts through other pages.

C. Castillo et al. [13].

- Task: given news published on Twitter, determine newsworthy topics, and check the credibility after selecting the newsworthy topics.
- Model: data labeling has been collected using Amazon Mechanical Turk, where users manually distinguish between newsworthy and non-newsworthy events on one hand and credible and non-credible news on the other hand. Training the supervised classifier was performed using features related to the message, the user, the topic, and the propagation of the post, with a best-feature selection process to reduce the list to 15 features.
- Evaluation: the supervised classifier achieved an accuracy of 86%, while propagation features proved to be one of the most relevant for assessing credibility.

5.3.3 Web Articles. Claims found in open domain settings like web articles don't abide by a unified format, as was the case for the Q&A forums, and social networks. In this case, the focus is generally reduced to the textual content of the claim, and the reliability of its source, regardless of how it is propagated.

K. Popat et al. [48].

- Task: assessing the credibility of textual claims from all kinds of web sources, including news websites, and social media. Without making any assumptions about the structure of the claim or the community where it was found.
- Model: a distant supervision model assess the credibility of claims using two types of features. Language stylistic features of the reporting articles, including assertive verbs, factive verbs, report verbs, hedges, implicative words, discourse markers, and lexicon of subjectivity and bias. Then it computes the normalized frequency of their occurrences, in order to construct a feature vector for each article. On the other hand, it captures the reliability of the web sources of the articles, using PageRank which considers the number and quality of links to and from the website, and AlexaRank which measures the page views and visitors of the website. The model is trained with the labels True and False attached to each claim, using an L1-regularized logistic regression model.
- Evaluation: to assess their performance, these models have been tested on two datasets: Snopes and Wikipedia. For each claim, 30 articles have been collected from the Google search engine. On the Snopes dataset, the model achieved 71.96% overall accuracy, 75.43% true accuracy, and 70.77% fake accuracy.

K. Popat et al. [49].

- Task: Similarly to the previous task [48].
- Model: The former distant supervision model was updated with a stance determination mode. Assuming that each claim has a reporting article, and each article is extracted from a web source, after finding the reporting article, they determined their stance by generating snippets of up to four sentences, calculating overlap between the claim and the snippets, and removing the ones with an overlap of less than a specified threshold. Afterward, they compute the stance probability of each snippet using a stance classifier and combine that value with the overlap score in order to rank the snippets based on the combined score and select the top-ranked ones to average their final stance score. Furthermore, unlike the distant supervision models that learn the features of the articles, sources, and claims separately. K. Popat et al. proposed an alternative: a probabilistic graphical model that consists of a conditional random field. This model captures the mutual interaction among the sources, articles, claims, and their credibility labels in tuples called cliques, where, each clique is supported by a set of associated feature functions with a weight vector.
- Evaluation: the distant supervision with stance achieved 81.39% overall accuracy, 83.21% true accuracy, and 80.78% false accuracy. While the CRF model achieved a better overall accuracy of 84.02%, but a lower 71.26% true accuracy, and 88.74% false accuracy.

5.4 Evidence Oriented Methods

This category of methods represents the core of this systematic review. In order to detect the scientific papers that match the claim studied, it is necessary to employ models that can calculate similarities between textual documents. Evidence-oriented methods generally apply neural network models. These methods put less focus on the surface of the claim, and more focus on learning the similarity between the candidate pieces of evidence and the claim studied. Regardless of the model utilized, some of these methods produce a prediction on the stance of the claim, but others aim to explain the rationale behind their prediction.

5.4.1 Stance Detection. Most of the stance detection methods came as proposed solutions to the Sci-Fact task. Other methods that didn't solve the Sci-Fact still use similar models. The only difference is that they resort to the whole web for evidence, and not specifically scientific research papers.

K. Popat et al. [50].

- Task: the task consists of predicting the credibility of a claim based on four elements: the claim, its source, an evidence article, and its source reporting article. Unlike K. Popat's previous works [49][48] that relied on manually crafted features and lexicon, and also other works like Wang's [68] and Rashkin's [53] where they trained neural networks on datasets of labeled claims, without resorting to external evidence. This method considers both the evidence from external sources and the representation of the claim itself.
- Model: a tuple of a claim, its source, the evidence article, article source, and the credibility label of the claim represent a training instance. The input claim and the input article are represented as a list of word or token embeddings. The sources of the claim and the article are represented by a multidimensional embedding vector. On the one hand, The articles' embeddings are fed to a bidirectional long short-term memory network. It generates a hidden vector for each token at each timestep. Bidirectional LSTMs are improved compared to standard LSTMS by capturing both previous and future timesteps through two hidden states. The past and future information are finally concatenated together

to form the final hidden vector. On the other hand, input claims are represented as the average of word embeddings, this value is concatenated with each article token. Then an activation function is applied to this concatenation to transform it into a claim-specific representation of each article token. A softmax activation follows this to calculate the attention score of each word of the article based on its relevance to the context of the claim. Finally, the weighted average of the product of the hidden vectors of the article and the attention scores of the words is calculated. Then it is combined with the embeddings of the sources of the claim and the article through two connected layers and a final softmax activation layer. The score achieved would be for one article. After reading a number of articles, we get the overall credibility score by averaging the sum of the scores achieved per article.

- Evaluation: Distant supervision with rich lexicons and feature engineering outperforms DeClarE in the Snopes dataset [49] on all scores. However, DeClare scores the best when tested on the Politifact[1], NewsTrust[41], and SemEval[15] datasets.

R. Pradeep et al. [51].

- Task: This method came as a proposed solution for the Sci-Fact task. This task consists of taking as input a scientific claim and a corpus of abstracts. The system must predict the gold evidence abstracts that support or refute the claim, identify up to 3 rationale sentences that justify the abstract prediction, and correctly predict the label of the claim.
- Model: VERT5ERINI starts with abstract retrieval. This consists of a two-stage pipeline starting with using the bag of words to rank abstracts, then re-ranking them using a T5 sequence-to-sequence model based on Nogueira et al.[46] that scores the abstracts based on relevance. Second, rationale selection, which is performed similarly to abstract retrieval using a T5 model, while filtering out the sentence with a probability under 0.999%. Negative samples are taken as non-evidence sentences from relevant abstracts. Finally, label prediction. Given the selected rational sentences, the T5 model assigns the probability of each of the three possible labels, and the highest probability label is selected
- Evaluation: At the time of publishing the paper, training the model on both the training and development set performed better than all submitted projects on sentence-level evaluation F1 by scoring 63.4, while scoring 66.95 on abstract-level evaluation.

X. Zeng et al. [72].

- Task: This method came as a proposed solution for the Sci-Fact task.
- Model: QMUL-SDS starts with abstract retrieval. The top 30 abstracts are ranked based on TF-IDF similarity ranking, then a trained BioBERT text classifier [28] filters the candidate abstracts based on their titles. Rationale selection is solved similarly. A BioBERT classifier [28] identifies the rationale sentences, making binary predictions about whether the sentences contain the rationale. For label prediction, the system switches to a RoBERTa model [33] to predict the labels. In the first stage, the supporting and contradicting labels have been merged into an "enough info" label. The model performs a binary classification on abstracts as either holding evidence or not. In the second stage, the contradicting and not enough info labels are merged into "not support", and a binary classification is performed again. The second classifier is only used if the first classifier outputs enough info.
- Evaluation: the QMUL-SDS achieves an F1 score of 55.35 on the sentence level and 58.38 on the abstract level.

X. Li et al. [29].

- Task: This method came as a proposed solution for the Sci-Fact task.
- Model: Paragraph-Joint starts with abstract retrieval. BioSentVec which is the biomedical version of Sent2Vec is leveraged for an efficient sentence-level similarity computation between the abstract embeddings and claim embeddings. The top 3 similar abstracts are selected as candidates. Then a joint rationale selection and stance prediction model is employed. First, a compact paragraph encoding is performed by concatenating the claim and the whole abstract paragraph as a single sequence and encoding it using a RoBERTa model [33]. Sentences are contextually represented as weights with respect to the claim and paragraph. Then the probability of each sentence being a rationale is computed. Two variants are used for stance prediction: a kernel graph attention network [35], and simple sentence-level attention. Training the model was performed with a negative sampling of irrelevant abstracts with high lexical similarity to the gold abstracts.
- Evaluation: The paragraph-joint model performs best on abstract-level evaluation with an F1 score of 67.16, while the sentence-level evaluation scored 60.94 on F1.

D. Wadden et al. [66].

- Task: This method came as a proposed solution for the Sci-Fact task.
- Model: MultiVers employs the VERT5ERINI retrieval system to retrieve the candidate abstracts, seeing that it achieved state-of-the-art performance on the SciFact dataset. A shared encoding of the claim, the abstract, and its title is then produced. The documents considered sometimes exceed the standard 512 token limit. Therefore, the long former model by I. Beltagy et al. [10] is used as the encoder. Every sentence from the abstract is predicted to be a rationale or not through a binary classification model. Similarly, the label is predicted through a three-way classification model over the rationale sentence encoding. The training of the model is performed on a combination of general domain annotations and weakly labeled in-domain data. Afterward, the model is finetuned on the SciFact dataset.
- Evaluation: The MultiVers model achieves the highest F1 scores on the abstract level 72.5 and sentence level 67.2.

Z. Liu et al. [66].

- Task: This method came as a proposed solution for the Sci-Fact task.
- Model: SCIKGAT starts with retrieving the top 100 abstracts using TF-IDF. Then a BERT model takes the representation of the concatenation of the claim, each abstract, and its title, and calculates the relevance between the claim and the abstracts. The abstracts are reranked accordingly, and only the top 3 abstracts are kept. Rationale selection is performed similarly. the relevance between the claim and the sentences is calculated and the highest-scoring sentences are reserved to form the retrieved evidence. For fact verification, the system switches to a kernel graph attention network [35] to predict the probability of the claim label based on each piece of evidence. The model was trained on the COVID-Fact dataset to incorporate useful medical words.
- Evaluation: Running the full model on the testing set achieves an F1 score of 50.48 on the sentence level and 58.33 on the abstract level

G. Karadzhov et al. [23].

- Task: fact-checking textual claims by retrieving external evidence from the entire web.
- Model: the claim is converted to a query of 5 to 10 tokens, excluding part-of-the-speech words that are not named entities, verbs, nouns, or adjectives. Then the query is applied to a search engine to retrieve relevant web pages. Each web page is represented by a snippet. The similarity between the claim and the snippets is calculated, and the highest similarity snippet is taken. In parallel, the web page is split into sentence triplets which embeddings are calculated as the average of the embeddings of its words. The features selected include the embedding of the claim, the embedding of the best-scoring snippet, and the embedding of best scoring sentence triplet from the web page. Two classifiers are built, a neural network classifier and a support vector machine classifier. In the neural network, the vectors of the aforementioned embeddings are concatenated and connected to a hidden layer. Then the final hidden layer of the neural network combines the task-specific embedding of the claim together with all the above evidence about it. Finally, a softmax output unit leads to classifying the claim as true or false. In the support vector machine, all the features are applied using a radial basis function kernel to classify the claim. Features here are calculated only by averaging, without a bi-LSTM. A joint model combining both the NN and the SVM that augments the input of the SVM by using values of units in the hidden layer of the NN was also suggested.
- Evaluation: The models have been tested on the rumor detection dataset by J. Ma et al. [38] and the cQA-QA-2016 dataset [43]. NN outperforms SVM on both datasets while using the joint model outperforms both models individually.

Y. Nie et al. [45].

- Task: This task consists of taking a textual claim and extracting the matching evidence documents and sentences in order to verify the credibility of the claim.
- Model: the Neural Semantic Matching Network contains four layers, an encoding layer which consists of a bidirectional LSTM that encodes the sequence tokens, an alignment layer that aligns two input sequences based on the encodings of their tokens, a matching layer that matches the two aligned sequences via a recurrent network, and an output layer that compresses the sequences into two vectors, and maps them along with their absolute difference to the final output.
  This network is applied under three homogeneous neural semantic matching models for each of the three following tasks: document retrieval, sentence selection, and claim verification. In document retrieval, keyword matching scores were calculated to retrieve the top k documents. Model training was performed by classifying disambiguative documents providing ground truth evidence as positive and the other disambiguative documents as negative.
  In sentence selection, they retrieved the set of evidential sentences from the retrieved documents by conducting the semantic matching model between each sentence and the claim.
  In claim verification, they conducted the semantic matching model with the input of the claim and the concatenation of evidential sentences in addition to three token-level features in order to conclude the label of the claim. In the last two tasks, model training was performed using the FEVER training set [63].
- Evaluation: the final model achieved a score of 66.14 on FEVER.

J. Ma et al. [37].

- Task: the task is similar to the aforementioned NSMN method, however, it differs from it by training to learn evidence representation. It also builds on the DeClare method by using evidential sentence embeddings.
- Model: The model introduced is an end-to-end hierarchical attention network for claim verification.
  It contains a sentence encoder that converts the claim and sentences to a multidimensional vector of word embeddings.
  A coherence-based attention layer, which is a model pre-trained using a pairwise training strategy. It takes embeddings of evidential sentences and feeds them to a coherence model that captures the coherence between the sentences regarding the set as a whole and the other sentences individually. We get a multidimensional vector where each element presents the attention weight of each sentence, after which the overall coherence of the sentence is calculated through the weighted sum of all sentences, and then concatenated with the original embedding of the sentence for a richer representation.
  An entailment-based attention layer captures whether the evidence infers a given claim. The model is pre-trained using the Stanford Natural Language Inference dataset [11]. The entailment model takes the previously calculated embeddings in addition to the embeddings of the claim and transforms them into a multidimensional vector that

presents the entailment score for each sentence.
Finally, an output layer predicts the verdict using a probability distribution over the veracity classes.
- Evaluation: using the Snopes dataset, the model achieves a true precision of 0.637 and false precision of 0.899, which scores higher than the CNN-based model [68], the LSTM-based model [53], the SVM model [62] and DeClarE [50]. On Politifact, it achieved a true precision of 0.495 and a false precision of 0.629. It also records a Fever score of 0.571.

A. Soleimani et al. [60].

- Task: This method came as a proposed solution for the FEVER fact extraction and verification challenge.
- Model: The approach is a pipeline composed of document retrieval, sentence retrieval, and claim verification.
  For document retrieval, BERT adopts the same component of the highest scoring team at the time of the study, UKP-Athene [19], achieving more than 93% for document recall. They used MediaWiki AP to inspect the Wikipedia database for noun phrases matching the claims.
  For sentence retrieval, the goal is to select the top 5 nearest sentences from the retrieved documents. The BERT model takes as input the potential evidence sentence and the claim. Token representations of the two sentences are all fed into an embedding layer for token, sentence, and positional embedding. Then a transformer encoder of 12 layers as a base mode, and 24 layers as a large mode is applied. Finally, the classification layer predicts the label of the sentence retrieved. The sentence retrieval is applied on a pointwise approach, which adds a cross-entropy layer after the model, and a pairwise approach where two models are applied to a negative sentence and a positive sentence along with the studied claim. Their predicted scores are then compared and fed to a final layer where a RankNet loss or a Hinge loss function is applied.
  For claim verification, the top 5 evidence sentences retrieved are compared. The final result would be NotEnoughInfo, unless there is supporting evidence for the claim supported. If there is one evidence sentence retrieved that rejects the claim, the label classification would be Refuted.
- Evaluation: the large mode combined with the pointwise approach achieved the second highest fever score of 69.66% on the test dataset after DREAM [73] by W. Zhong et al.

W. Zhong et al. [73].

- Task: This method came as a proposed solution to the FEVER challenge.
- Model: Similarly to the aforementioned evidence-focused methods, the pipeline is composed of document retrieval, sentence selection, and evidence retrieval. For document retrieval, keyword matching is applied, then NSMN[45] is applied to assign higher rankings to documents with disambiguation titles, and the top 10 documents are selected. For sentence selection, pre-trained models like XLNet[69] are applied to calculate the similarity of a claim to the list of candidate sentences from the extracted documents. The top 5 sentences are selected.
  For evidence retrieval, a graph-based reasoning method is employed. This approach understands the semantic structure of the pieces of evidence by representing them as a graph. Constructing a graph given a group of sentences could be performed using named entity recognition and open information extraction. In this paper, they applied semantic role labeling SRL [12], which operates the following way: each sentence is parsed to a tuple. In each tuple, the verb, argument, location, and temporal elements represent nodes of the graph. Every two nodes are linked through an edge. Nodes across different tuples are also linked through edges to capture relationships between multiple sentences. The same concept is applied to the claim sentence and retrieved evidence sentences. They are fed to a transformer-based pre-trained model in order to capture their contextual word representations along with their graph distances. They are then averaged at the node level and are fed to multi-layer graph convolutional networks that aggregate the representations of the nodes and their neighbors into a graph. This is performed separately for evidence-based graphs and claim-based graphs. Finally, the node representations of the two graphs are aligned in a graph attention network before making the final prediction.
- Evaluation: the DREAM approach achieved the highest FEVER score of 70.60%.

5.4.2 Justification Production. Justification production methods are not completely different from stance prediction methods. However, the former usually requires a knowledge base or a list of rules to produce justifications. This section presents a few justification production methods that can be applied to fact-checking claims by relying on evidence from scientific papers in addition to a knowledge base.

M. Gad-Elrab et al. [16].

- Task: The purpose of this study is to generate explanations for fact-checking a claim which is represented as a rule. Given a knowledge graph, a text corpus, a set of rules, and a global parameter that ensures termination of rewriting. The aim is to detect concise, close-to-the-query, and reliable explanations.
- Model: The algorithm defines the following functions:
  A textspot function that receives as an input a set of textual documents, and a query which is a relation between two entities. It returns the set of strings related to the query.
  A bind function that receives as input a text and a knowledge graph and a query that is a relation between a known entity and an unknown entity. It returns the set of substitutions answering the query preferably from the KG and possibly from the text. This function iterates over the detected substitutions and includes them in the final set of explanations.
  A rewrite function that receives as input a set of defined atoms, a set of rules, and a query that is a relation between a known entity and an unknown entity. It returns the union between the set of atoms and the substitution of the body of the query from the set of rules. This function rewrites the

input query, if it's not already a fact, using the termination parameter.
The final output is a set of explanations, which is a union between the set of rules and the facts involving all entities and relations that can be extracted using textspot. The set of explanations can serve as input features in order to establish the truth value of the fact.
- Evaluation: Compared to prior methods, the ExFaKT algorithm achieved relatively higher accuracy in the Politicians benchmark dataset [42] against Truthfinder [70] and LSC [49], but had lower recall against LSC in both the Politicians benchmark dataset and the DBpedia-based dataset [59].

P. Atanasova et al. [7].

- Task: The aim of the study is to implement a model that fact-checks claims and learns to extract explanations similar to the given human justifications, by taking as input the claim and a list of ruling comments based on its context.
- Model: This study makes use of a PolitiFact-based dataset called LIAR-PLUS. This dataset already contains veracity justifications for each of the 12,836 textual claims. The model used for explanation extraction is based on DistilBERT [57] which is a reduced version of BERT. The model produces the contextual embeddings of the claim and the ruling sentences and feeds them to a pre-trained function that selects the top 4 sentences that constitute the most similar explanations to the human justification.
The veracity prediction is produced similarly, by feeding the claim and the ruling sentences to the pre-trained DistilBERT model that produces the contextual embeddings and applies a trained function that predicts the veracity of the claim.
Veracity production can be combined with veracity explanation generation through a joint model that parallelizes the functions of the aforementioned models through a function that learns to predict both the veracity explanation and veracity label of the claim.
- Evaluation: The joint model ranks better in coverage and overall studied criteria, while the individual explanation generation model achieved better ROUGE F1 scores, and was ranked higher in non-redundancy and non-contradiction.

N. Ahmadi et al. [4].

- Task: Given a claim, a knowledge graph, and web sources, this approach aims to produce explanations for the verdict predicted on the claim.
- Model: This approach resorts to a knowledge graph to discover positive rules and to the web to discover negative rules relevant to the claim. The head variables of the extracted rules are replaced by the claim values. Then they generate and unify all the evidence triples from the KG and the web that have a valid substitution to the body of the rules. After collecting the relevant rules, an inference method adopted by the probabilistic answer set program LPMLN [27] is applied to predict the decision of the claim. This method relies on assigning weights to every rule to calculate the probability of the model based on the normalized weight of its involved rules. The goal is to produce explanations that consist of the union of substitutions for the body of the relevant rules.
- Evaluation: Datasets have been extracted from DBPedia [59], based on predicates consisting of spouse, deathplace, vicepresident, and almaMater. The method consisting of the LPMLN inference model, coupled with evidence collected from the web and the KG scored the best with an average F1 score on all predicates of 0.81.

6 RESULTS

The results of this study are presented as a summary of all the methods discussed and a comparison of their main attributes. Based on the focus of the method, I showcase these attributes through Table 2 which addresses the claim-oriented methods, and Table 3 which addresses the evidence-oriented methods.

6.1 Claim Oriented Methods

Table 2 presents five claim-oriented methods. The methods for the claims found in Q&A forums came as proposed solutions for the 8th task of the SemEval challenge. We can notice that both methods utilized identical verdict prediction models and training and testing datasets. However, the increase in accuracy is mainly due to the update in the training dataset. The update consisted of only including questions with at least 10 answers. Given that the list of answers is considered a feature to measure the truthfulness of the claim, it makes sense to populate the dataset with more textual context to improve the accuracy. This speaks of the importance of having high-quality large-scale training datasets.
The social media method [13] used a supervised classifier since social networks are full of metadata revolving around the post, the user, the propagation of the post, and its topic. A long list of features could be engineered using the surface of the claim.
The web methods are also similar in terms of scope, training, and modeling. The second work by Popat [49], however, achieves high average accuracy by considering the reporting articles of the claims and their stances. It also adds a conditional random field that captures the mutual interaction between the claims, their reporting articles, and their source reliabilities. Although this might not strictly lead to better performance as the achieved accuracy of the True claims was lower than the first work's accuracy of True claims [48], the average accuracy was still better due to the CRF model being more biased towards the false claims.

6.2 Evidence Oriented Methods

Table 3 presents fourteen evidence-oriented methods. Looking at the Challenge and Dataset columns, it is clear that these methods can be compared by separating them into three categories. Methods based on the Sci-Fact dataset, methods based on the FEVER dataset, and methods based on the DBPedia dataset. Apart from these, we also find two studies related to the Snopes dataset. First, Popat's DeClare [50] still underperforms compared to the distant supervision and CRF method by the same author [49]. Second, Karadzhov's system trained on Snopes and cQA-QA-2016 also underperforms as compared to the claim-oriented Q&A methods on the updated

Task			Claim			Verdict			Dataset		
Scope	Author	Challenge	Type	Detection	Matching	Type	Prediction	Justification	Training	Testing	Evaluation
Q&A	Mihaylova[40]	SemEval 8	Textual	-	-	Binary	SVM	-	CQA-QL-2016	CQA-QL-2016	83% Acc
Q&A	Nokav[44]	SemEval 8	Textual	-	-	Binary	SVM	-	CQA-QL-2016 Update	CQA-QL-2016 Update	91.81% Acc
Social Media	Castillo[13]	-	Textual	Twitter	-	Binary	Supervised Classifier	-	Mechanical Turk	Mechanical Turk	86.01% Acc
Web	Popat[48]	-	Textual	Google	-	Multi-label	Distant Supervision	-	Snopes	Snopes	71.96% Acc
Web	Popat[49]	-	Textual	Google	-	Multi-label	Distant Supervision + CRF	-	Snopes	Snopes	84.02% Acc

Table 2: Claim Oriented Methods Attributes

Author	Challenge	Claim	Evidence			Verdict			Dataset		
		Type	Type	Retrieval	Source	Type	Prediction	Justification	Training	Testing	Evaluation
Popat[50]	-	Textual	Unstructured	Bi-LSTM	Web	Scale + Multi-label	Attention Weights	-	Snopes	Snopes	78.96% True Acc
Pradeep[51]	Sci-Fact	Textual	Unstructured	T5	Abstracts	Multi-label	T5	-	Sci-Fact	Sci-Fact	63.4% A 66.95% S F1
Zeng[72]	Sci-Fact	Textual	Unstructured	TF-IDF + BioBERT	Abstracts	Multi-label	RoBERTa	-	Sci-Fact	Sci-Fact	58.38% A 55.35% S F1
Li[29]	Sci-Fact	Textual	Unstructured	BioSentVec + RoBERTa	Abstracts	Multi-label	KGAT	-	Sci-Fact	Sci-Fact	67.16% A 60.94% S F1
Wadden[66]	Sci-Fact	Textual	Unstructured	VERT5ERINI	Abstracts	Multi-label	Classification	-	General + Sci-Fact	Sci-Fact	72.5% A 67.2% S F1
Liu[34]	Sci-Fact	Textual	Unstructured	TF-IDF + BERT	Abstracts	Multi-label	KGAT	-	COVID-Fact + Sci-Fact	Sci-Fact	50.48% A 58.33% S F1
Karadzhov[23]	-	Textual	Unstructured	TF-IDF	Google + Bing	Binary	NN + SVM	-	Snopes + cQA-QA-2016	cQA-QA-2016	72.7% F1 + Acc
Nie[45]	-	Textual	Unstructured	Semantic matching	Wikipedia	Multi-label	Semantic matching	-	FEVER	FEVER	66.14% FEVER
Ma[37]	-	Textual	Unstructured	Coherence attention	Wikipedia	Multi-label	Entailment attention	-	SNLI	FEVER	57.1% FEVER
Soleimani[60]	FEVER	Textual	Unstructured	Multi-sentence entailment	Wikipedia	Multi-label	Classification	-	FEVER	FEVER	69.66% FEVER
Zhong[73]	FEVER	Textual	Unstructured	Graph-based reasoning	Wikipedia	Multi-label	KGAT	-	FEVER	FEVER	70.6% FEVER
Gad-Elrab[16]	-	Triple	Graph + Unstructured	Textspot + Bind functions	Wikipedia	Multi-label	Human annotation	Rewrite function	DBpedia	DBpedia	93% Acc
Atanasova[7]	-	Textual	Structured	-	LIAR-PLUS	Multi-label	DistilBERT	DistilBERT	LIAR-PLUS	LIAR-PLUS	31.58% ROUGE-L
Ahmadi[4]	-	Triple	Graph + Unstructured	Text mining	Bing	Multi-label	LPMLN	Rule mining	DBpedia	DBpedia	81% F1

Table 3: Evidence Oriented Methods Attributes

rumor detection dataset.

When it comes to Sci-Fact, there are five selected methods that were proposed as solutions to the challenge. Retrieval of abstracts and rationale sentences was operated with a variety of methods. The most common ones were TF-IDF similarity and BERT-based models. Knowing that the abstract-level evaluation considers the predicted label of the claim as well, we can assume that it is a better measure of veracity prediction models, while sentence-level evaluation is a better measure of evidence retrieval models. The lowest-scoring models at the sentence level were the models that used TF-IDF similarity, while the highest-scoring models by Pradeep [51] and Wadden [66] both used an identical system consisting of a T5 sequence-to-sequence model to retrieve both the relevant abstracts and the rationale sentences. The highest-scoring models at the abstract level were not very different. The MultiVers model by Wadden[66] came first again. While VERT5ERINI came third, handing the second place to SCIKGAT by Li et al. [34]. Looking at the training datasets, it seems that MultiVers was the only method that pre-trained its model on a combination of general domain annotations and weakly labeled in-domain data before training on Sci-Fact. The weak supervision might have also contributed to the increase in retrieval precision. The paper also emphasized the importance of the long former encoder that allows verifying claims against longer contexts, like full scientific papers.

When it comes to FEVER, four methods were highlighted. Unlike the Sci-Fact methods, they use Wikipedia articles as a source of retrieving evidence rather than abstracts from research papers. The evaluation as well is done by testing on the FEVER dataset which leads to measuring a unified FEVER score. All methods trained their models on the FEVER dataset, apart from the lowest-scoring method by Ma [37] that used the Stanford Natural Language Inference dataset for training. The two highest-scoring methods by Soleimani [60] and Zhong [73] used KGAT and classification similar to the two highest-scoring SciFact methods on the abstract level evaluation. However, the graph structure that re-defines the relative distances of semantically related words performs better in selecting the relevant evidence sentences to the claims and predicts their credibility more accurately.

Finally, two of the justification production methods trained and tested their models on the DBpedia dataset. While Gad-Elrab [16] and Ahmadi [4] used DBpedia, Atanasova [7] uses LIAR-PLUS which is an updated version of PolitiFact. These methods all had different variables, in terms of claim, evidence, verdict, and evaluation, therefore, it is difficult to objectively compare their attributes. However, the two DBpedia-based methods are similar in their types of claims and evidence. They also both employed the web as a backup for the knowledge graph in order to generate substitutions of the claim triples and produce explanations through a union of substitutions. However, Ahmadi's work [4] adds the probabilistic answer set program LPMLN in order to calculate the probability of the truthfulness of the claim, while Gad-Elrab [16] presents the explanations as an input feature that can be used automatically or manually to annotate the claim.

7 DISCUSSION

The discussion aims to combine the learnings achieved from the results section and attempt to provide a suggestion for a generally suitable and applicable method for the task at hand: fact-checking a textual statement by resorting to scientific research papers.

As the result section shows, it is complicated to compare these methods only through conducting research and without producing any experimental data. This is due to many reasons: the variety of evaluation methods, types of claims, types of evidence collected, sources of evidence, training datasets, testing datasets, and most importantly, the bias of the authors in evaluating their proper methods.

The aforementioned comparison conducted in the results section, however, proves that some methods do share some of these attributes, which facilitates revealing some insights. A number of the discussed methods came as proposed solutions for the same challenge, and thereby also have a uniform evaluation system.

Comparing the claim-oriented methods revealed that well-populated datasets that were verified against large textual context usually train models to perform better. While the Sci-Fact submissions demonstrated the importance of pre-training on general domain annotations and weakly labeled data. For this, a suggestion would be to pre-train the model at hand at a large and general-purpose dataset like FEVER which contains 185,444 claims, or MultiFC which contains 34,918 claims. Then using SciFact-Open instead of SciFact because the former extended the verification to a large corpus of 500K research abstracts. Sci-Fact Open also proved to be a difficult dataset since the top 5 systems developed for Sci-Fact underperformed by 15% to 30% on F1.

Assuming that the textual claim is not to be found in an open-domain setting where the metadata of the claim or the source could be utilized, only a combination of evidence-oriented techniques can be employed. The highest scoring Sci-Fact system originally proposed by VERT5ERINI [51] ranked the relevant scientific papers through a bag of words-based system, then reranked them using a T5 sequence to sequence model [46]. The same method could be employed to detect the most relevant sentences. A longformer encoder [10] would also reduce the complexity of the model and allow taking processing of a large number of tokens, thereby, considering large parts of scientific research papers, instead of only taking their abstracts. After which, a simple classification could output the veracity prediction.

Another option to explore is a graph-based reasoning approach [73]. By contextualizing the word representations of the claim and the word representations of the evidence with the distances between their nodes in two separate evidence-based and claim-based graphs. It is possible to predict the veracity of the claim by aligning the two graphs in a graph attention network. Using a knowledge graph attention network could also be a good veracity prediction technique, as it was often utilized in the top-scoring methods for both the FEVER and the Sci-Fact challenges.

8 CONCLUSION

In this work, I present a systematic review of the state-of-the-art methods that can be utilized in the context of fact-checking textual statements through evidence collected from scientific research papers. After giving an overview of the fact-checking pipeline, I present the methods under two main categories: claim-oriented methods, and evidence-oriented methods. Although it is complicated to compare and evaluate methods with no experimental setup, the main findings consisted of combining the learnings achieved and giving a few suggestions of techniques to follow in order to perform fact-checking on textual statements using scientific research. Future work on this topic consists of putting to the test the suggestions about the choice of training datasets, evidence retrieval models, and verdict prediction methods, and implementing a method in the same scope of the systematic review. This implementation would be part of my thesis project where I aim to scope down further the context of the fact-checking system to analyze a specific group of claims or a specific source of evidence. This will enable me to confirm or falsify the hypotheses made in the results and discussion section of this study.

REFERENCES

[1] [n.d.]. https://www.politifact.com/

[2] 2016. Fact- checking. https://reporterslab.org/fact-checking/

[3] 2018. Fact checking. https://en.wikipedia.org/wiki/Fact_checking

[4] Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mohammed Saeed. 2019. Explainable Fact Checking with Probabilistic Answer Set Programming. https://doi.org/10.36370/tto.2019.15

[5] Pepa Atanasova, Ivan Koychev, Lluís Màrquez, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. A Context-Aware Approach for Detecting Check-Worthy Claims in Political Debates.

[6] Pepa Atanasova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James R. Glass. 2019. Automatic Fact-Checking Using Context and Discourse Information. *CoRR* abs/1908.01328 (2019). arXiv:1908.01328 http://arxiv.org/abs/1908.01328

[7] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating Fact Checking Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7352–7364. https://doi.org/10.18653/v1/2020.acl-main.656

[8] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Lima, Casper Hansen, Christian Hansen, and Jakob Simonsen. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims.

[9] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. 3606–3611. https://doi.org/10.18653/v1/D19-1371

[10] Iz Beltagy, Matthew Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer.

[11] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 632–642. https://doi.org/10.18653/v1/D15-1075

[12] Xavier Carreras and Lluís Màrquez. 2004. Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*. Association for Computational Linguistics, Boston, Massachusetts, USA, 89–97. https://aclanthology.org/W04-2412

[13] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. *Proceedings of the 20th International Conference on World Wide Web*, 675–684. https://doi.org/10.1145/1963405.1963500

[14] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1870–1879. https://doi.org/10.18653/v1/P17-1171

[15] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 69–76. https://doi.org/10.18653/v1/S17-2006

[16] Mohamed Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. 2019. ExFaKT: A Framework for Explaining Facts over Knowledge Graphs and Text. 87–95. https://doi.org/10.1145/3289600.3290996

[17] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2021. A Survey on Automated Fact-Checking. (08 2021).

[18] Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A Richly Annotated Corpus for Different Tasks in Automated Fact-Checking. 493–503. https://doi.org/10.18653/v1/K19-1046

[19] Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-Athene: Multi-Sentence Textual Entailment for Claim Verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics, Brussels, Belgium, 103–108. https://doi.org/10.18653/v1/W18-5516

[20] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. 1803–1812. https://doi.org/10.1145/3097983.3098131

[21] Naeemul Hassan, Anil Nayak, Vikas Sable, Chengkai Li, Mark Tremayne, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, and Aaditya Kulkarni. 2017. ClaimBuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*

10 (08 2017), 1945–1948. https://doi.org/10.14778/3137765.3137815
[22] Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, and Smaranda Muresan. 2020. DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
[23] Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. Fully Automated Fact Checking Using External Sources. 344–353. https://doi.org/10.26615/978-954-452-049-6_046
[24] Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2018. Towards Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. *CoRR* abs/1809.08193 (2018). arXiv:1809.08193 http://arxiv.org/abs/1809.08193
[25] Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking: A Survey. 5430–5443. https://doi.org/10.18653/v1/2020.coling-main.474
[26] Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking for Public Health Claims.
[27] Joohyung Lee and Yi Wang. 2016. Weighted Rules under the Stable Model Semantics *(KR'16)*. AAAI Press, 145–154.
[28] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics (Oxford, England)* 36 (09 2019). https://doi.org/10.1093/bioinformatics/btz682
[29] Xiangci Li, Gully Burns, and Nanyun Peng. 2020. A Paragraph-level Multi-task Learning Model for Scientific Fact-Verification.
[30] Xian Li, Xin Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. 2015. Truth Finding on the Deep Web: Is the Problem Solved? *Proceedings of the VLDB Endowment* 6 (03 2015).
[31] Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2016. A Survey on Truth Discovery. *SIGKDD Explor. Newsl.* 17, 2 (feb 2016), 1–16. https://doi.org/10.1145/2897350.2897352
[32] Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders.
[33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://doi.org/10.48550/ARXIV.1907.11692
[34] Zhenghao Liu, Chenyan Xiong, Zhuyun Dai, Sun Si, Maosong Sun, and Zhiyuan Liu. 2020. Adapting Open Domain Fact Extraction and Verification to COVID-FACT through In-Domain Language Modeling. 2395–2400. https://doi.org/10.18653/v1/2020.findings-emnlp.216
[35] Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained Fact Verification with Kernel Graph Attention Network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7342–7351. https://doi.org/10.18653/v1/2020.acl-main.655
[36] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S. Weld. [n.d.]. S2ORC: The semantic scholar open research corpus. https://aclanthology.org/2020.acl-main.447/
[37] Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019. Sentence-Level Evidence Embedding for Claim Verification with Hierarchical Attention Networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2561–2571. https://doi.org/10.18653/v1/P19-1244
[38] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting Rumors from Microblogs with Recurrent Neural Networks *(IJCAI'16)*. AAAI Press, 3818–3824.
[39] Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. 2015. Finding Opinion Manipulation Trolls in News Community Forums. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Beijing, China, 310–314. https://doi.org/10.18653/v1/K15-1032
[40] Tsvetomila Mihaylova, Georgi Karadzhov, Pepa Atanasova, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. 2019. SemEval-2019 Task 8: Fact Checking in Community Question Answering Forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 860–869. https://doi.org/10.18653/v1/S19-2149
[41] Subhabrata Mukherjee and Gerhard Weikum. 2015. Leveraging Joint Interactions for Credibility Analysis in News Communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (Melbourne, Australia) *(CIKM '15)*. Association for Computing Machinery, New York, NY, USA, 353–362. https://doi.org/10.1145/2806416.2806537
[42] Ndapandula Nakashole and Tom M. Mitchell. 2014. Language-Aware Truth Assessment of Fact Candidates. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, 1009–1019. https://doi.org/10.3115/v1/P14-1095
[43] Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. SemEval-2016 Task 3: Community Question Answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, 525–545. https://doi.org/10.18653/v1/S16-1083
[44] Preslav Nakov, Tsvetomila Mihaylova, Lluís Màrquez, Yashkumar Shiroya, and Ivan Koychev. 2017. Do Not Trust the Trolls: Predicting Credibility in Community Question Answering Forums. 551–560. https://doi.org/10.26615/978-954-452-049-6_072
[45] Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining Fact Extraction and Verification with Neural Semantic Matching Networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence* (Honolulu, Hawaii, USA) *(AAAI'19/IAAI'19/EAAI'19)*. AAAI Press, Article 842, 8 pages. https://doi.org/10.1609/aaai.v33i01.33016859
[46] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. https://doi.org/10.48550/ARXIV.2003.06713
[47] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A Decomposable Attention Model for Natural Language Inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2249–2255. https://doi.org/10.18653/v1/D16-1244
[48] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility Assessment of Textual Claims on the Web. 2173–2178. https://doi.org/10.1145/2983323.2983661
[49] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media. https://doi.org/10.1145/3041021.3055133
[50] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning.
[51] Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2020. Scientific Claim Verification with VERT5ERINI.
[52] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.
[53] Hannah Rashkin, Eunsol Choi, Jin Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. 2931–2937. https://doi.org/10.18653/v1/D17-1317
[54] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.
[55] Alejandro Romero and Jonathan Nelson. 2022. Will the world clean up 'information pollution' in 2022? https://www.weforum.org/agenda/2022/03/misinformation-cybercrime-covid-pandemic/
[56] Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic.
[57] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.
[58] Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based Fact-Checking of Health-related Claims. 3499–3512. https://doi.org/10.18653/v1/2021.findings-emnlp.297
[59] Prashant Shiralkar, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. 2017. Finding Streams in Knowledge Graphs to Support Fact Checking. In *2017 IEEE International Conference on Data Mining (ICDM)*. 859–864. https://doi.org/10.1109/ICDM.2017.105
[60] Amir Soleimani, Christof Monz, and Marcel Worring. 2019. BERT for Evidence Retrieval and Claim Verification. *CoRR* abs/1910.02655 (2019). arXiv:1910.02655 http://arxiv.org/abs/1910.02655
[61] James Thorne and Andreas Vlachos. 2017. An Extensible Framework for Verification of Numerical Claims. https://doi.org/10.18653/v1/E17-3010
[62] James Thorne and Andreas Vlachos. 2018. Automated Fact Checking: Task Formulations, Methods and Future Directions. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3346–3359. https://aclanthology.org/C18-1283
[63] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 809–819. https://doi.org/10.18653/v1/N18-1074
[64] Jacopo Urbani. 2022. Web Data Processing Systems. https://studiegids.vu.nl/EN/courses/2021-2022/XM_40020#/
[65] David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-open: Towards open-domain scientific claim verification. https://arxiv.org/abs/2210.13777

[66] David Wadden, Kyle Lo, Lucy Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. 61–76. https://doi.org/10.18653/v1/2022.findings-naacl.6

[67] David Wadden, Kyle Lo, Lucy Wang, Shanchuan Lin, Madeleine Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims.

[68] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, 422–426. https://doi.org/10.18653/v1/P17-2067

[69] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. Curran Associates Inc., Red Hook, NY, USA.

[70] Xiaoxin Yin, Jiawei Han, and Philip Yu. 2008. Truth Discovery with Multiple Conflicting Information Providers on the Web. *Knowledge and Data Engineering, IEEE Transactions on* 20 (07 2008), 796 – 808. https://doi.org/10.1109/TKDE.2007.190745

[71] Xia Zeng, Amani Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass* 15 (10 2021). https://doi.org/10.1111/lnc3.12438

[72] Xia Zeng and Arkaitz Zubiaga. 2021. QMUL-SDS at SCIVER: Step-by-Step Binary Classification for Scientific Claim Verification.

[73] Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning Over Semantic-Level Graph for Fact Checking. 6170–6180. https://doi.org/10.18653/v1/2020.acl-main.549